Journal of Econometrics, 2008, 142(1), pp. 1-27

Nonlinearity, Nonstationarity, and Spurious Forecasts*

Vadim Marmer†

August 2004 (Revised: May 2007)

Abstract

Implications of nonlinearity, nonstationarity and misspecification are considered from a forecasting perspective. Our model allows for small departures from the martingale difference sequence hypothesis by including a nonlinear component, formulated as a general, integrable transformation of the I(1) predictor. We assume that the true generating mechanism is unknown to the econometrician and he is therefore forced to use some approximating functions. It is shown that in this framework the linear regression techniques lead to spurious forecasts. Improvements of the forecast accuracy are possible with properly chosen nonlinear transformations of the predictor. The paper derives the limiting distribution of the forecasts' MSE. In the case of square integrable approximants, it depends on the L_2-distance between the nonlinear component and approximating function. Optimal forecasts are available for a given class of approximants.

*I am grateful to Don Andrews, Peter Phillips, Yuichi Kitamura, Patrik Guggenberger, Erik Hjalmarsson, Joann Jasiak, Alex Maynard, Randi Pintoff, Kevin Song, two anonymous referees, seminar participants at Yale University, Bar-Ilan University, Ben-Gurion University, Hebrew University of Jerusalem, Rice University, Tel Aviv University, SUNY at Albany, UBC, University of Haifa, Université de Montéal, University of Western Ontario, Board of Governors of the Federal Reserve System, as well as participants of CIRANO - CIREQ Conference on Forecasting in Macroeconomics and Finance, 22nd Canadian Econometrics Study Group Conference, and 2006 Far Eastern Meeting of the Econometric Society for many helpful comments and discussions. I thank the Cowles Foundation for support under a Cowles Prize.

†Department of Economics, University of British Columbia, 997 - 1873 East Mall, Vancouver, BC V6T 1Z1, Canada. E-mail: vmarmer@interchange.ubc.ca.

1 Introduction

Nonlinear models are extensively used in econometrics (for a description and analysis of various nonlinear models see, for example, Granger and Teräsvirta (1993)). The theoretical foundation for estimation of nonlinear, nonstationary models has been developed fairly recently. Park and Phillips (1999) derived asymptotic results for the sums of nonlinear transformations of integrated time series. They considered three classes of nonlinear functions: integrable, homogeneous and exponential. They show, for example, that partial sums of integrable functions that have a non-zero Lebesgue measure converge in distribution to local times of the Brownian motion. Their results have been applied to various nonlinear econometric models. Chang et al. (2001) considered nonlinear regression with separably additive regression functions. Chang and Park (2003) considered nonstationary index models, which extend switching regressions to the stochastic trends framework. Hu and Phillips (2004) studied nonstationary discrete choice models. Kasparis (2004) considered effects of functional form misspecification on estimation, when the true and estimated models involve nonlinearity and nonstationarity. He focused on convergence of estimators to some pseudo-true values and detection of functional form misspecification. Hong and Phillips (2005) develop a linearity test of cointegrating relations.

An attractive feature of nonlinear models is flexibility that allows one to model relationships between nonstationary and seemingly stationary variables. A linear regression requires the dependent variable to have the same order of integration as the right-hand side of the regression equation. However, it is known that nonlinear transformations can change the memory properties of a process. Thus, contrary to linear regressions, properly chosen nonlinear functions can link in a single equation variables that appear to have different orders of integrations. Nonlinear functions that can be used to model relationships between seemingly stationary and persistent variables include Lebesgue integrable functions and asymptotically homogeneous functions of degree zero (the distribution-like functions). For example, Chang and Park (2003) modelled nonstationary switching behavior by adding a distribution type function of the I(1) variable to a noise process.

There are many situations in economics that may require one to relate variables of different orders of integration. A typical example is the predictive regressions literature in empirical finance, which studies stock returns predictability. In a predictive

regression, stock returns are regressed on lagged values of various financial and economic variables such as the dividend-price ratio, earnings-price ratio or interest rates. While most researchers agree that stock market returns are I(0), predictors such as dividend-price ratio appear to have a stochastic trend component.

In predictive regressions, predictability is usually concluded on the basis of t-tests for slope coefficients. Often, it is implicitly assumed that non-zero in-sample correlations found between regressors and stock returns can be used for construction of out-of-sample forecasts. Many papers report statistically significant slope estimates (see, for example, Fama (1991) and Cochrane (1997) for surveys of the literature). Despite the collected empirical evidence on in-sample relations between stock returns and predictors, the out-of-sample predictability is still a controversial issue. Goyal and Welch (2003, 2004) report that performance of out-of-sample forecasts based on linear regression methods can be rather poor, while Campbell and Thompson (2004) argue that there exists small but economically meaningful out-of-sample predictive power, once restrictions on the coefficients and forecasts are imposed.

The results in this paper imply that significant regression slopes do not necessarily indicate usefulness of the linear regression as a forecasting equation. Our model allows for small departures from the martingale difference sequence (MDS) hypothesis by including an additive nonlinear component, formulated as a general, integrable transformation of the predictor, which is assumed to be I(1). In this model, the signal coming from the nonlinear component is very weak relative to the noise, as implied by the properties of integrable functions and I(1) variables. An integrable function approaches zero at a fast rate as the absolute value of its argument increases. At the same time, a unit root process usually takes on very large negative or positive values. As a result, the signal coming from the predictor (the nonlinear component) is relatively strong only during rare events, when the unit root process visits the neighborhood of zero. Such a generating mechanism provides for predictability only in the extremely short run, which in the stock market example corresponds to a situation where some relevant information may escape the attention of market participants only for very short periods of time. The process modelled in this paper is oposite to that of Kilian and Taylor (2003); they describe nonlinear mean-reversion, however, combined with long-run predictability (they discuss forecasting of exchange rates).

It is natural to assume that the true data generating process (DGP) involving nonlinear dependency is unknown to the econometrician, and he is therefore forced

to use some approximating functions. Furthermore, the class of approximants used by the econometrician does not necessarily include the true function. This paper illustrates by the means of analytical asymptotic results that such a combination of nonstationarity, nonlinearity and misspecification leads to the results often seen in the predictive regressions literature. Consider for example a linear regression, which is the most popular approximating function. We show that, in this case, commonly used diagnostic tools tend to indicate predictive power despite the fact that estimated regression slopes converge to zero in probability. Moreover, we show that the out-of-sample forecasts constructed from a predictive regression asymptotically have the same mean squared error (MSE) as that of constant forecasts equal to the simple historic average of the dependent variable. Hence, spurious forecasts occur: diagnostic tools may indicate usefulness of the model, while, in fact, equivalent or better forecasts may be obtained, if one completely ignores the information contained in the predictor.

The predictability in our model is very limited due to the nature of the DGP. Nevertheless, out-of-sample forecasting accuracy can be improved by using square integrable approximating functions instead of historic averages or linear regressions. We derive the limiting distribution of the out-of-sample MSE. In the case of square integrable approximants, it depends on the L_2-distance between the nonlinear component and approximating function. We show that, for a given class of square integrable approximating functions, one can obtain the best forecasts in the MSE sense.

In this paper, we consider only the case of a single predictor, since this is the situation usually studied in the predictive regressions literature. However, the framework can be easily generalized to the case of multiple predictors with an additively separable regression function.

The paper is organized as follows. Section 2 introduces the model and reviews the asymptotic theory of nonlinear functions of integrated processes. In Section 3, we consider the class of forecasts constructed as polynomials in the predictor. This class contains predictive regressions as a special case. Section 4 discusses forecasting with integrable approximants. Section 5 presents the predictability tests based on square integrable transformations. We report simulation results in Section 6, and Section 7 provides an empirical example. Section 8 concludes. All proofs are given in the Appendix.

2 Definitions and preliminary results

In this paper, we consider a nonstationary nonlinear model described by the following assumptions.

Assumption 2.1. **(a)** $y_t = \mu^* + f(z_{t-1}) + u_t$, where μ^* is an unknown constant, $f : R \to R$ is an unknown nonlinear function, and $\{(u_t, z_t) : t = 1, ..., n\}$ are random variables such that

(b) $z_t = z_{t-1} + C(L)\varepsilon_t$, where $C(L) = \sum_{k=0}^{\infty} c_k L^k$, and L is the lag operator;

(c) $\{((u_t, \varepsilon_t), \mathcal{F}_t) : t \geq 1\}$, where $\mathcal{F}_t = \sigma(\{(u_s, \varepsilon_s) : s \leq t\})$, is a stationary and ergodic MDS;

(d) $z_0 = 0$.

Assumption 2.1(c) implies that the error process u_t cannot be predicted from its lagged values and the lagged values of the predictor z_t. The assumption $z_0 = 0$ in (d) is imposed for convinience. The results of this paper essentially will not change if we allow z_0 to be of a more general form $z_0 = n^{1/2}c$ for some constant c. We make the following additional assumptions concerning the innovations process $\{(u_t, \varepsilon_t) : t \geq 1\}$.

Assumption 2.2. **(a)** $E\left((u_t, \varepsilon_t)'(u_t, \varepsilon_t) | \mathcal{F}_{t-1}\right) = \Sigma > 0$ a.s. for all t.

(b) $\sup_{t \geq 1} E\left(|u_t|^h | \mathcal{F}_{t-1}\right) < \infty$ for some $h > 2$.

(c) $\{\varepsilon_t : t \geq 1\}$ are iid with $E|\varepsilon_1|^q < \infty$ for some $q > 4$.

(d) $C(1) \neq 0$ and $\sum_{k=0}^{\infty} k|c_k| < \infty$.

(e) The distribution of ε_1 is absolutely continuous with respect to the Lebesgue measure and $\left|Ee^{ik\varepsilon_1}\right| = o\left(k^{-\beta}\right)$ for some $\beta > 0$.

It is assumed that f belongs to the class of integrable regular (I-regular) functions denoted by $\mathcal{I}$.

Definition 2.1. *A function $\varphi : R \to R$ is said to belong to the class of I-regular functions, denoted by $\mathcal{I}$, if*

(a) *φ is Lebesgue integrable.*

(b) *$\int_{-\infty}^{\infty} \varphi(x)\, dx \neq 0$.*

(c) *For some constants $c > 0$ and $k > 6/(q-2)$ with $q > 4$ given in Assumption 2.2(c), $|\varphi(x) - \varphi(y)| < c|x - y|^k$.*

The class of I-regular functions and the smoothness condition in (c) were introduced in Park and Phillips (1999). We make the following assumption concerning f.

Assumption 2.3. f and $f^2 \in \mathcal{I}$.

Functions that satisfy Assumption 2.3 must have tails decreasing at sufficiently fast rate. At the same time, since z_t is an integrated process, it takes on very large negative or positive values most of the time. As a result, $f(z_{t-1})$ is usually close to zero, except for the periods when z_{t-1} visits the neighborhood of zero. The integrated variable z_{t-1} becomes a useful predictor for y_t only on such rare occasions. Furthermore, sample paths of $f(z_{t-1}) + u_t$ appear to be similar to the sample paths of the noise process u_t (see Section 6 for an example).

Assumption 2.1 define a predictive model. An optimal forecast in the MSE sense is given by

$$\widehat{y}_t(\widehat{\mu}) = \widehat{\mu} + f(z_{t-1}),$$

where $\widehat{\mu}$ is a Least Squares (LS) estimator of μ. However, the optimal forecast is infeasible if f is unknown to the econometrician. In this case, the econometrician is forced to use some approximating functions instead of f. We assume that the class of approximating functions considered by the econometrician does not include the true function f, and the forecasts are constructed using a misspecified model. We denote the approximating function by $g(\cdot, \theta)$, where θ is a vector of constants. The value of θ is chosen by the econometrician. In this paper, we consider two alternatives for $g(x, \theta)$. The first is a polynomial function in x, which includes the predictive regression ($g(x, \theta) = \theta x$) and constant forecasts ($g(x, \theta) = 0$ for all $x \in R$) as special cases. Since the true DGP involves a square integrable function, the second type of approximating functions considered here consists of square integrable functions in x.

The results of this paper are derived under so called fixed forecasting scheme (see, for example, West and McCracken (1998)). The econometrician observes the data $\{(y_t, z_{t-1}) : t = 1, ..., n_1\}$. His objective is to construct one period ahead forecasts of y_t for periods $\{n_1 + 1, ..., n\}$ using the actual values of z_{t-1}, which are observed before y_t is realized. While the results of this paper continue to hold true for short forecasting horizons of more than one period, we do not consider long forecasting horizons, since our model captures only very short-run predictability. The forecasting equation is defined as

$$\widehat{y}_t(\mu, \theta) = \mu + g(z_{t-1}, \theta), \tag{2.1}$$

where $\widehat{y}_t$ is the predicted value of y_t, the function g is chosen by the econometrician, and μ and θ are scalar and vector constants respectively. We assume that μ and θ

are estimated from the training sample $\{(y_t, z_{t-1}) : t = 1, ..., n_1\}$. Let $\left(\widehat{\mu}, \widehat{\theta}'\right)$ be an estimator of (μ, θ'). The econometrician uses the estimated version of equation (2.1) in order to construct forecasts for the observations in the forecasting sample, which consists of observations $n_1 + 1$ through n. We assume that $n_1 = [nr]$, where $r \in (0, 1)$ and fixed, and $[\cdot]$ denotes the integer part.

Following the approach of Diebold and Mariano (1995), in this paper forecasting models are evaluated according to their out-of-sample performance relative to that of a benchmark model. Specifically, we assume that the forecasts are evaluated with a quadratic loss function:

$$Q_n\left(\widehat{\mu}, \widehat{\theta}\right) = (n - n_1)^{-1} \sum_{t=n_1+1}^{n} \left(y_t - \widehat{\mu} - g\left(z_{t-1}, \widehat{\theta}\right)\right)^2 - (n - n_1)^{-1} \sum_{t=n_1+1}^{n} u_t^2.$$

The first term in the above expression is the sample MSE of the series of forecasts

$$\left\{\widehat{y}_t\left(\widehat{\mu}, \widehat{\theta}\right) : t = n_1 + 1, ..., n\right\}.$$

The second term is the MSE of the infeasible forecasts

$$\{\mu^* + f(z_{t-1}) : t = n_1 + 1, ..., n\}.$$

The second term does not depend on the choice of the forecasting function or the values of μ and θ, and, therefore, minimization of $Q_n(\mu, \theta)$ is achieved only through minimization of the first component.

Write

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_{zu} \\ \sigma_{zu} & \sigma_u^2 \end{pmatrix}.$$

Under Assumptions 2.1(b)-(d) and 2.2, in large samples, the distribution of a process

$$\left(n^{-1/2} z_{[nr]}, n^{-1/2} \sum_{t=1}^{[nr]} u_t\right)$$

can be approximated by the distribution of a two-dimensional Brownian motion with the covariance matrix

$$\Omega = \begin{pmatrix} \omega_z^2 & \omega_{zu} \\ \omega_{zu} & \sigma_u^2 \end{pmatrix},$$

where $\omega_z^2 = C(1)^2\sigma_z^2$ is the long-run variance of the innovations of z_t, and ω_{zu} is the long-run covariance of u_t and the innovations of $z_t : \omega_{zu} = \sum_{h=1}^{\infty} E(\varepsilon_h u_1)$. Note that the correlations between u_t and the lagged values of ε_t are equal to zero due to the MDS assumption.

We assume that the parameters in equation (2.1) are estimated by LS, which leads to expressions of the form $\sum_t f(z_{t-1})$. Asymptotically, partial sums of integrable transformations of I(1) variables are approximated by local times of a Brownian motion. The local time (L) of the Brownian motion B at $x \in R$ is defined as

$$L(t,x) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t 1_{(x-\varepsilon, x+\varepsilon)}(B(s))\, ds,$$

where 1_A is an indicator function of the set $A \subset R$. The local time measures the amount of time that the Brownian motion B spends in the neighborhood of the point x (see, for example, Chung and Williams (1990) for an introduction to local time). The following result, due to Park and Phillips (1999, 2001), is the foundation of the subsequent discussion.

Lemma 2.1. Let $\varphi_h : R \to R$ be a homogeneous function of degree $h > 0$, $\varphi_I \in \mathcal{I}$, and $\varphi_{2,I}$ be such that $\varphi_{2,I}^2 \in \mathcal{I}$. Let w_u and w_z be two independent standard Brownian motions. Define $\ell(t) = \omega_z^{-1} L_z(t,0)$, where L_z is the local time of w_z. If Assumptions 2.1(b)-(d) and 2.2 hold, then for fixed $0 < s < r < 1$ the following results hold jointly:

(a) $\left(n^{-1/2}\sum_{t=[ns]}^{[nr]} \varepsilon_t, n^{-1/2}\sum_{t=[ns]}^{[nr]} u_t\right) \to_d \left(\int_s^r db(t), \int_s^r du(t)\right)$, where $(b,u)' = \Omega^{1/2}(w_z, w_u)'$.

(b) $n^{-1-h/2}\sum_{t=[ns]}^{[nr]} \varphi_h(z_{t-1}) \to_d \int_s^r \varphi_h(b(t))\, dt$.

(c) $n^{-1/2-h/2}\sum_{t=[ns]}^{[nr]} \varphi_h(z_{t-1})\, u_t \to_d \int_s^r \varphi_h(b(t))\, du(t)$.

(d) $n^{-1/2}\sum_{t=[ns]}^{[nr]} \varphi_I(z_{t-1}) \to_d \left(\int_{-\infty}^{\infty} \varphi_I(x)\, dx\right) \int_s^r d\ell(t)$.

(e) $n^{-1/4}\sum_{t=[ns]}^{[nr]} \varphi_{2,I}(z_{t-1})\, u_t \to_d \sigma_u \left(\int_{-\infty}^{\infty} \varphi_{2,I}^2(x)\, dx \int_s^r d\ell(t)\right)^{1/2} W(1)$, where W is a Brownian motion independent of (w_z, w_u).

The initial condition on z_0 has an effect on the asymptotic results in Lemma 2.1(d) and (e) through the local time process. If one generalizes Assumption 2.1(d) as $z_0 = n^{1/2}c$, the local time at zero process would be replaced with the local time at $-c$ (see, for example, Phillips et al. (2007)).

3 Forecasting with polynomials

In this section, we consider forecasts constructed as polynomials in lagged values of the predictor. In this case,

$$g\left(z_{t-1},\theta\right)=\sum_{j=1}^{p}\theta_j z_{t-1}^j, \tag{3.2}$$

where p is a positive integer chosen by the econometrician. For $p=1$, equation (3.2) reduces to a simple linear regression considered in the predictive regression literature. In addition to Assumption 2.3, we assume that f satisfies the following condition.

Assumption 3.4. $f(x)x^p \in \mathcal{I}$.

It is assumed that θ is estimated by LS using the training sample. Collecting the powers of z_t , we define:

$$\begin{aligned} Z_t &= \left(z_t,\ldots,z_t^p\right)', \\ \underline{Z}_t &= Z_t - [nr]^{-1}\sum_{s=1}^{[nr]} Z_{s-1}. \end{aligned} \tag{3.3}$$

The LS estimator of μ and θ in (3.2) is given by

$$\begin{aligned} \widehat{\theta}_n &= \left(\sum_{t=1}^{[nr]}\underline{Z}_{t-1}\underline{Z}_{t-1}'\right)^{-1}\sum_{t=1}^{[nr]}\underline{Z}_{t-1}y_t, \\ \widehat{\mu}_n &= [nr]^{-1}\sum_{t=1}^{[nr]} y_t - [nr]^{-1}\sum_{t=1}^{[nr]} Z_{t-1}'\widehat{\theta}_n. \end{aligned} \tag{3.4}$$

Suppose that the econometrician draws a conclusion regarding the predictability of y_t from a test based on the usual Wald statistic for θ:

$$\begin{aligned} F_n &= \widehat{\theta}_n'\left(\sum_{t=1}^{[nr]}\underline{Z}_{t-1}\underline{Z}_{t-1}'\right)\widehat{\theta}_n/\widehat{\sigma}_{u,n}^2, \text{ where} \\ \widehat{\sigma}_{u,n}^2 &= [nr]^{-1}\sum_{t=1}^{[nr]}\left(y_t-\widehat{\mu}_n-Z_{t-1}'\widehat{\theta}_n\right)^2. \end{aligned} \tag{3.5}$$

We compare the forecasts constructed according to model (3.2) with a baseline

model that assumes no predictability:

$$\widehat{y}_t(\mu_0, 0) = \mu_0. \tag{3.6}$$

Equation (3.6) is a particular case of a polynomial forecasting function. It depends on a single parameter μ_0, which is estimated by the average of y_t in the training sample:

$$\widehat{\mu}_{0,n} = [nr]^{-1} \sum_{t=1}^{[nr]} y_t.$$

Similarly to (3.3), we define

$$\begin{aligned} B(t) &= (b(t), \ldots, b(t)^p)', \\ \underline{B}(t) &= B(t) - r^{-1} \int_0^r B(s)ds, \end{aligned}$$

where the Brownian motion b as in Lemma 2.1. The following theorem describes the asymptotic behavior of the estimators $\widehat{\theta}_n$ and $\widehat{\mu}_n$, test statistic F_n, and loss function Q_n. Let $D_n = \text{diag}(n, \ldots, n^p)$.

Theorem 3.1. *Under Assumptions 2.1-3.4,*

(a) $n^{1/2} D_n^{1/2} \widehat{\theta}_n \to_d \Psi$, *where*

$$\begin{aligned} \Psi &= \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds \right)^{-1} \int_0^r \underline{B}(s)du(s) \\ &\quad - r^{-1}\ell(r) \left(\int_{-\infty}^{\infty} f(x)dx \right) \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds \right)^{-1} \int_0^r B(s)ds. \end{aligned} \tag{3.7}$$

(b) $n^{1/2}(\widehat{\mu}_n - \mu^*) \to_d r^{-1} \left(\ell(r) \int_{-\infty}^{\infty} f(x)dx + u(r) - \Psi' \int_0^r B(s)ds \right)$.

(c) $F_n \to_d \left\| \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds \right)^{1/2} \Psi/\sigma_u \right\|^2$.

(d) *For* $p \geq 0$, $n^{1/2} Q_n \left(\widehat{\mu}_n, \widehat{\theta}_n \right) \to_d (1-r)^{-1} (\ell(1) - \ell(r)) \int_{-\infty}^{\infty} f^2(x)dx$.

(e) *For* $p > 0$, $n \left(Q_n \left(\widehat{\mu}_n, \widehat{\theta}_n \right) - Q_n \left(\widehat{\mu}_{0,n}, 0 \right) \right) \to_d \Delta$, *where the support of the random variable* Δ *includes the negative and positive parts of the real line.*

Part (a) of the theorem implies that $\widehat{\theta}_n$ converges in probability to zero. When rescaled by $n^{1/2} D_n^{1/2}$, in large samples, its distribution can be approximated by the distribution of the random variable defined in (3.7). The first term on the right-hand

side of (3.7) is the usual expression obtained in the limit, when one regresses an I(0) variable on I(1) regressors. This term has a mixed normal distribution when $\omega_{zx} = 0$. The second term in equation (3.7) comes from the nonlinear part of the DGP. It depends both on the integral of f over the entire real line and on the local time at zero of the limiting process of z_t. The second component gives the mean of the mixed normal distribution, when u_t and ε_t are uncorrelated at all lags and leads. Thus, due to nonlinear nonstationarity, the asymptotic distribution of $\widehat{\theta}_n$ is noncentral.

Noncentrality of the asymptotic distribution of $\widehat{\theta}_n$ translates to that of the F_n statistic. Despite the fact that $\widehat{\theta}_n$ converges to zero in probability, part (c) of the theorem implies that a test based on F_n tends to reject the hypothesis of no predictability (in the current context, the hypothesis of no predictive power is equivalent to $\theta = 0$). For example, consider the case $\omega_{zx} = 0$. In this case, F_n has a mixed noncentral χ^2_p distribution with the noncentrality parameter given by

$$\left(\frac{\int_{-\infty}^{\infty} f(x)dx}{\sigma_u}\right)^2 \left\| \left(r^{-1}\ell(r)\int_0^r \underline{B}(s)\underline{B}'(s)ds\right)^{-1/2} \int_0^r B(s)ds \right\|^2 .$$

Consequently, the test that rejects the null of no predictability when $F_n > \chi^2_{p,1-\alpha}$, where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$ quantile of a central χ^2_p distribution, rejects the null with probability greater than α. Actual rejection probabilities depend on the ratio of the integral of f to the standard deviation of the noise process u_t (the signal-to-noise ratio in this framework). Interestingly, the shape of the nonlinear function f has no effect on the rejection rates, since f appears in the expression for the noncentrality parameter only through its integral over the entire real line. Note, however, that F_n statistic is not diverging, a result related to that of Hong and Phillips (2005), Theorem 8. Higher than nominal rejection rates are due to noncentrality of the asymptotic distribution of F_n.

Lastly, parts (d) and (e) of the theorem shows that Q_n has the same limiting distribution regardless of the value of p. In particular, the baseline model ($p = 0$) asymptotically yields the same loss as a model with $p > 0$. Moreover, the asymptotic distribution of the loss function does not depend on the information contained in the predictor. Therefore, the inclusion of powers of the predictor in a forecasting equation does not improve the forecast accuracy. Part (e) of the theorem describes the asymptotic distribution of the difference of the loss functions for the polynomial

and baseline forecasting models. The expression for Δ is given in the Appendix. The support of the limiting distribution of the difference includes both positive and negative parts of the real line. In section 6, we show with the help of Monte Carlo simulations that this difference tends to be positive in finite samples and, therefore, that the baseline model dominates polynomials in the MSE sense.

4 Forecasting with integrable functions

The previous section illustrates that the polynomials can be poor predictors, when the DGP involves nonstationarity and nonlinearities of a certain type. In fact, a simple average can dominate a polynomial forecasting model in terms of the MSE. The reason for this lies in the global nature of the LS approximation in the current framework. Consider minimization of the L_2-distance between $f(x)$ and the approximating function $g(x,\theta)$:

$$\inf_{\theta\in\Theta}\int_{-\infty}^{\infty}\left(f(x)-g(x,\theta)\right)^2 dx, \tag{4.8}$$

where $\Theta \subset R^p$ is a compact set. Suppose that $g(x,\theta)$ is unbounded and diverges to $\pm\infty$ as $x \to \pm\infty$, which is true for polynomials. In this case, a solution to (4.8) requires a choice of θ such that $g(x,\theta) = 0$ almost always. This illustrates why, in the previous section, the limit of the loss function Q_n is proportional to $\int_{-\infty}^{\infty} f^2(x)dx$ and does not depend on the information contained in the predictor. The situation changes if one uses square integrable approximating functions instead of polynomials. If $g(x,\theta)$ is square integrable, then a non-trivial solution to (4.8) exists, which leads to improvements in forecast accuracy. This occurs since

$$\int_{-\infty}^{\infty} f^2(x)dx \geq \inf_{\theta\in\Theta}\int_{-\infty}^{\infty}\left(f(x)-g(x,\theta)\right)^2 dx,$$

whenever there exists $\widetilde{\theta} \in \Theta$ such that $g\left(x,\widetilde{\theta}\right)$ is equal to zero almost everywhere.

In this section, we consider forecasting with square integrable (with respect to x) functions $g(x,\theta)$ in the forecasting equation (2.1). The function g has to be nonlinear in x due to the integrability assumption; however, it may depend on θ in a linear or nonlinear way. We assume that the econometrician restricts θ to a compact subset of R^p, denoted by Θ. The dimension of θ is chosen by the econometrician together with

the functional form of g. Linear (in θ) forecasting functions are of greatest interest:

$$g(x,\theta) = \sum_{i=1}^{p} \theta_i \phi_i(x), \tag{4.9}$$

where ϕ_i's are integrable for $1 \leq i \leq p$. An example of a nonlinear function in θ is the class of extended rational polynomials (ERPs):

$$g(x,\theta) = \phi(x) \frac{\tau_0 + \tau_1 x + ... + \tau_p x^p}{1 + \beta_1 x + ... + \beta_p x^p}, \tag{4.10}$$

where ϕ is integrable and $\theta = \left(\tau_0, \tau_1 ..., \tau_p, \beta_1, ..., \beta_p\right)$. Functions of this type were used by Phillips (1983) for density approximation. The approximating functions considered in this paper are described by the following definition and assumption.

Definition 4.2. *A function $\varphi : R \times \Theta \to R$ is said to belong to the class of I-regular functions on Θ, denoted by $\mathcal{I}(\Theta)$, if*

(a) *for each $\theta_0 \in \Theta$, there exists a neghborhood N_0 of θ_0 and $T : R \to R$ bounded and integrable such that $|g(x,\theta) - g(x,\theta_0)| \leq \|\theta - \theta_0\| T(x)$ for all $\theta \in N_0$;*

(b) *for all $\theta \in \Theta$, $g(\cdot,\theta) \in \mathcal{I}$.*

The I-regular functions on a set were introduced in Park and Phillips (2001). We make the following assumption.

Assumption 4.5. (a) $\Theta \subset R^p$ is compact.

(b) g and $g^2 \in \mathcal{I}(\Theta)$.

The solution to the problem described in (4.8) depends on the choice of g. In some cases, such as (4.9) and (4.10), there exists a unique θ that solves (4.8). However, in general, multiple solutions may exist. Let Θ^* be the set of solutions to (4.8):

$$\begin{aligned} \Theta^* &= \left\{ \theta^* \in \Theta : \int_{-\infty}^{\infty} \left(f(x) - g(x,\theta^*)\right)^2 dx = M^* \right\}, \text{ where} \\ M^* &= \inf_{\theta \in \Theta} \int_{-\infty}^{\infty} \left(f(x) - g(x,\theta)\right)^2 dx. \end{aligned}$$

Note that different choices of g and p lead to different M^* and Θ^*. Any value $\theta^* \in \Theta^*$ can be treated as a pseudo-true value of θ for a given choice of g.

Similarly to Section 3, we assume that μ and θ are estimated by LS from the training sample (nonlinear LS, if g is nonlinear in θ). In the second step, the estimates of μ and θ are used to compute predicted values of y in the forecasting sample. Let Θ_n be the set of values of $\theta \in \Theta$ that solve the in-sample LS problem:

$$\min_{\theta \in \Theta, \mu} \sum_{t=1}^{[nr]} \left(y_t - \mu - g\left(z_{t-1}, \theta\right)\right)^2.$$

For each $\widehat{\theta}_n \in \Theta_n$, the corresponding estimator of μ is given by

$$\widehat{\mu}_n\left(\widehat{\theta}_n\right) = [nr]^{-1} \sum_{t=1}^{[nr]} \left(y_t - g\left(z_{t-1}, \widehat{\theta}_n\right)\right). \tag{4.11}$$

We define the distance between θ and the set $A \subset R^p$ as

$$d\left(\theta, A\right) = \inf_{a \in A} \left\|a - \theta\right\|.$$

The following result describes the behavior of the LS estimators of θ and μ and error function Q_n as the sample size approaches infinity.

Theorem 4.2. *Under Assumptions 2.1-2.3 and 4.5,*

(a) $\sup_{\widehat{\theta}_n \in \Theta_n} d\left(\widehat{\theta}_n, \Theta^*\right) \to_p 0.$

(b) $\sup_{\widehat{\theta}_n \in \Theta_n} \left|\widehat{\mu}_n\left(\widehat{\theta}_n\right) - \mu^*\right| \to_p 0.$

(c) $\sup_{\widehat{\theta}_n \in \Theta_n} n^{1/2} Q_n\left(\widehat{\mu}_n\left(\widehat{\theta}_n\right), \widehat{\theta}_n\right) \to_d \left(1 - r\right)^{-1} \left(\ell(1) - \ell(r)\right) M^*.$

Theorem 4.2(a) implies that $\widehat{\theta}_n$ converges in probability to the set of its pseudo-true values. It follows from part (b) of the Theorem that $\widehat{\mu}_n\left(\widehat{\theta}_n\right)$ is a consistent estimator of μ^*. Part (c) of the Theorem shows that, in the case of square integrable approximating functions, asymptotically Q_n is proportional to the least distance between the true nonlinear function and its approximant: $M^* = \inf_{\theta \in \Theta} \int_{-\infty}^{\infty} \left(f(x) - g\left(x, \theta\right)\right)^2 dx$. Thus, asymptotically one achieves the smallest out-of-sample MSE for a given class of functions g. Finally, comparison of the results of Theorem 3.1(d) and 4.2(c) implies that integrable functions yield an improvement in the forecasting accuracy over polynomials and the baseline model.

For certain choices of g, the solution to (4.8) is unique, i.e. $\Theta^* = \{\theta^*\}$. In this case, using the results of Jeganathan (2003), one can show that $\widehat{\theta}_n$ converges in

probability to θ^* at the rate of $n^{1/4}$, which is slower than usual. While its asymptotic distribution is mixed normal, the covariance matrix depends on functionals of the unknown function f. Further, $\widehat{\mu}_n$ has the usual $n^{1/2}$ rate of convergence to μ^*. Its limiting distribution is mixed normal and centered around

$$\ell(r) \int_{-\infty}^{\infty} \left(f\left(x\right) - g\left(x, \theta^*\right)\right) dx. \tag{4.12}$$

The results of Theorem 4.2 remain valid if one replaces $\widehat{\mu}_n$ with the average of y_t in the training sample, $\widehat{\mu}_{0,n}$. It may have an advantage over $\widehat{\mu}_n$, since the asymptotic distribution of the historic average is centered around

$$\ell(r) \int_{-\infty}^{\infty} f\left(x\right) dx. \tag{4.13}$$

While we have that $\int_{-\infty}^{\infty} \left(f\left(x\right) - g\left(x, \theta^*\right)\right) \partial g\left(x, \theta^*\right) / \partial \theta dx = 0$, nevertheless, using a simple average to estimate μ^* may result in smaller bias (4.13) than that of the LS estimator (4.12).

5 Testing predictability

This section presents testing procedures that rejects the null of no predictability only if the corresponding model possesses out-of-sample predictive power superior to that of the baseline model. It is convenient to consider the class of linear approximants defined by equation (4.9). In this case, Θ^* is singleton, and the hypothesis of interest is whether $\theta^* = 0$. The linear in parameters forecasting model is computationally simple. However, its advantages are not limited to computational convenience. For nonlinear in parameters approximants, some parameters may be unidentified under the null. For example, in the case of ERPs described in equation (4.10), under the null of no predictability $\tau_0^* = \tau_1^* = ... = \tau_p^* = 0$. However, the coefficients in the denominator $\beta = \left(\beta_1, ..., \beta_p\right)'$ are not identified under the null. Any value of β in some compact set would give asymptotically the same result.

5.1 In-sample test

In view of the results presented in the previous sections, we propose a modification of predictive regressions based on integrable transformations of the predictor. We consider a local to zero alternative DGP:

Assumption 5.6. $y_t = \mu + n^{-1/4} f(z_{t-1}) + u_t$.

Scaling by $n^{-1/4}$ instead of usual $n^{-1/2}$ in the alternative DGP follows from the convergence rates in Lemma 2.1(d) and (e). We make the following assumption regarding the basis functions ϕ_i in (4.9):

Assumption 5.7. ϕ_i and $\phi_i^2 \in \mathcal{I}$ for all $1 \leq i \leq p$.

Let $\Phi(z_t) = \left(\phi_1(z_t), ..., \phi_p(z_t)\right)'$. The LS estimator of θ is given by

$$\widehat{\theta}_n = \left(\sum_{t=1}^{n} \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right) \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right)' \right)^{-1} \sum_{t=1}^{n} \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right) y_t,$$

where $\overline{\Phi}_n = n^{-1} \sum_{t=1}^{n} \Phi_{t-1}$. The Wald test statistic for $H_0 : \theta^* = 0$ is defined as

$$T_n = \widehat{\theta}_n' \left(\sum_{t=1}^{n} \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right) \left(\Phi(z_{t-1}) - \overline{\Phi}_n\right)' \right) \widehat{\theta}_n / \widehat{\sigma}_{u,n}^2, \tag{5.14}$$

where $\widehat{\sigma}_{u,n}^2 = n^{-1} \sum_{t=1}^{n} \left(y_t - \widehat{\mu}_n - \Phi(z_{t-1})' \widehat{\theta}_n\right)^2$. In the case of non-local alternatives, the results in Section 4 imply that T_n diverges to plus infinity at the rate of $n^{1/2}$. Suppose one rejects the null hypothesis if $T_n > \chi^2_{p,1-\alpha}$, where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2_p distribution. The following theorem describes the asymptotic size of the test and its power against local alternatives.

Theorem 5.3. *Under Assumptions 2.1-2.3, 5.6 and 5.7, T_n has an asymptotically noncentral mixed χ^2_p distribution with the noncentrality parameter given by*

$$\left\| \frac{1}{\sigma_u} \left(\ell(1) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1/2} \int_{-\infty}^{\infty} f(x)\Phi(x) dx \right\|^2.$$

The theorem implies that the test based on T_n detects alternatives approaching

the null at the rate slower than $n^{-1/4}$. Next, note that

$$\theta^* = \left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1} \int_{-\infty}^{\infty} f(x)\Phi(x) dx.$$

The null hypothesis, $H_0 : \theta^* = 0$, is true if f and Φ are orthogonal, including the case of unrelated y_t and z_{t-1} ($f(x) = 0$ almost everywhere). In this case, asymptotically the test statistic T_n has a central χ^2_p distribution regardless of the value of the long-run covariance ω_{zu}. Furthermore, Theorems 4.2 and 5.3 together imply that, contrary to the test based on F_n, the test based on T_n does not tend to reject the null hypothesis of no predictive power, unless the forecasting model has a better out-of-sample fit than the baseline forecasting equation.

5.2 Out-of-sample tests

In this section, we consider the predictability tests based on out-of-sample statistics. Such a test has an advantage of explicitly testing the out-of-sample performance of a forecasting model. However, this approach requires splitting of the sample into training and forecasting subsamples. As a result, smaller number of observations is used in actual testing, which may lead to power loss comparing to the test discussed in the previous section (for a discussion of in-sample and out-of-sample predictability tests see Inoue and Kilian (2004)). We assume that the relative size of the training sample is $r \in (0, 1)$ and fixed by the econometrician. Again, we consider linear in parameters forecasting equation (4.9).

Let $Q_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right)$ and $Q_n\left(\widehat{\mu}_{0,n}, 0\right)$ be the loss functions corresponding to equation (4.9) and the baseline model respectively. The test statistic is defined as follows.

$$S_n = (1-r)\, n \left(Q_n\left(\widehat{\mu}_{0,n}, 0\right) - Q_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right) \right) / \widehat{\sigma}^2_{u,[nr]}.$$

If z_{t-1} has predictive power over y_t, as defined by Assumptions 2.1(a) and 2.3, then it follows from Theorems 3.1(d) and 4.2(c) that asymptotically $S_n/n^{1/2}$ is proportional to $\int_{-\infty}^{\infty} f^2(x)dx - \int_{-\infty}^{\infty} \left(f(x) - g\left(x, \theta^*\right)\right)^2 dx \geq 0$. This difference is strictly positive, unless the base functions ϕ_i in (4.9) are orthogonal to f. Consequently, in such a situation, S_n diverges to infinity at the rate of $n^{1/2}$. Thus, the significance level α test based on S_n is to reject the null hypothesis of no predictability, for the forecasting

model given by Φ, when

$$S_n > c_\alpha,$$

where c_α is a positive constant chosen so that the test has significance level α. The following theorem provides the asymptotic distribution of S_n under the $1/n^{1/4}$ local alternatives defined by Assumption 5.6.

Theorem 5.4. *Under Assumptions 2.1-2.3, 5.6 and 5.7, as $n \to \infty$, S_n converges in distribution to*

$$\left\| \frac{d(r)\ell(r)}{\sigma_u} \left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1/2} \int_{-\infty}^{\infty} f(x)\Phi(x)dx + W_2 \right\|^2 - \|d(r)W_1 - W_2\|^2,$$

where $d(r) = \left((\ell(1) - \ell(r))/\ell(r)\right)^{1/2}$, and W_1, W_2 are $N(0, I_p)$ random vectors independent from ℓ and each other.

Under the null, θ^* and, consequently, $\int_{-\infty}^{\infty} f(x)\Phi(x)dx$ are equal to zero, and the asymptotic distribution of S_n is described by the distribution of

$$\|W_2\|^2 - \|d(r)W_1 - W_2\|^2.$$

Power of the test depends on the magnitude of

$$\left\| \frac{d(r)\ell(r)}{\sigma_u} \left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1/2} \int_{-\infty}^{\infty} f(x)\Phi(x)dx \right\|^2.$$

For example, the test has poor power properties if the vector of base functions $\Phi(x)$ is close to being orthogonal to the true function f. Power of the test also decreases with the variance of the unpredictable component u_t.

In the definition of $d(r)$, ℓ stands for the local time of the limit process of $z_t/n^{1/2}$. Hence, the critical values cannot be tabulated and must be calculated on a case by case basis, given $\{z_t\}$. First, one has to obtain an estimate of $d(r)$. A natural estimator is $\widehat{d}_{hn}(r) = \left(\sum_{t=[nr]+1}^{n} h(z_t) / \sum_{t=1}^{[nr]} h(z_t) \right)^{1/2}$, where $h \in \mathcal{I}$, for example, the standard normal density function. Next, one simulates two independent $N(0, I_p)$ random vectors, say $W_{k,1}$ and $W_{k,2}$, and calculates $S_{hn,k} = \left\| \widehat{d}_{hn}(r)W_{k,1} - W_{k,2} \right\|^2 - \|W_{k,2}\|^2$. For large n, the distribution of $S_{hn,k}$ approximates the asymptotic distribution of S_n un-

der the null. Therefore, one can repeat this for $k = 1, \ldots, K$, and estimate c_α by the $1-\alpha$ sample quantile of $\{S_{hn,k} : k = 1, \ldots, K\}$. As we mention in the discussion after Lemma 2.1, if one assumes that $z_0 = n^{1/2}c$, then, in the asymptotic distribution of S_n, the local time at zero process will be replaced with local time at $-c$. However, in practice, estimation of c is not required, since for any value of c the statistic $\sum_t h(z_t)$ converges to its corresponding local time process.

6 Monte Carlo results

The theoretical results of the previous sections suggest that nonlinear processes described by Assumptions 2.1-2.3 may be confused with an MDS. In this case, usual linear regression methods lead to spurious forecasts, while integrable approximants provide better out-of-sample fit. This section presents a series of Monte Carlo experiments motivated by these findings. First, we would like to illustrate similarities between an MDS and a process generated according to the model in Assumptions 2.1-2.3. In this and the next section, we use ϕ to denote the standard normal density function.

Figure 1 shows a typical sample path of a process generated according to Assumptions 2.1-2.3 with $f(x) = 10\phi(x)$, the independent standard normal errors $\{u_t\}$, and random walk $\{z_t\}$ with the standard normal increments independent of $\{u_t\}$. The random walk was initialized at zero. The top graph plots the errors $\{u_t\}$, the graph in the middle shows the sample path of the nonlinear component $\{f(z_t)\}$, and the graph at the bottom shows the sum of the two components. The figure shows that the signal generated by the nonlinear part is strong relative to the noise only during the first 15 periods. After that, the sum of the nonlinear component and the noise cannot be distinguished from an MDS.

The results in Section 3 suggest that a test based on the usual linear regression methods (F_n) is likely to indicate predictive power, if Assumptions 2.1-2.3 provide a good approximation of the true DGP. It is important to see rejection rates in finite samples. We simulate the data according to the following equations:

$$\begin{aligned} y_t &= af(z_{t-1}) + u_t, \\ (u_t, \Delta z_t)' &\sim \text{iid N}(0, I_2), \\ z_0 &= 0, \end{aligned}$$

where the constant a allows one to vary the signal-to-noise ratio. We consider the following choices of f :

$$\begin{aligned} f_1(x) &= 1\{0 < x < 1\}, \\ f_2(x) &= (1 - 0.5x)1\{0 < x < 2\}, \\ f_3(x) &= x^2 1\left\{0 < x < 3^{1/3}\right\}, \\ f_4(x) &= 2\phi(x + 0.25) - \phi(x - 0.75). \end{aligned}$$

The four functions have different graphs, however, they all have the same Lebesgue measure 1. Thus, according to the results in Section 3, all four functions should provide similar rejection rates. This is due to the fact that the asymptotic distribution of F_n depends on the integral of f and not on the shape of the function.

We construct F_n using the LS estimates of $(\theta_1, ..., \theta_p)$ in the forecasting model below:

$$\widehat{y}_t(\mu, \theta) = \mu + \sum_{i=1}^{p} \theta_i z_{t-1}^i.$$

Table 1 reports the simulated rejection rates for the sample size of 100 observations. The number of simulations is 1,000, the nominal size is 5%, $a \in \{1, 2, 4, 8\}$, and $p \in \{1, 2, 3, 4\}$. Table 1 shows that the actual rejection rates are higher than the nominal 5%. For example, in the case of a usual predictive regression ($p = 1$), the rejection rates are around 20% for $a = 2$, and they exceed 50% for most of the models when $a = 8$. Thus, the econometrician is likely to conclude that the polynomial forecasting function has predictive power. Simulations also confirm that the shape of f does not have an effect on the rejection rates, since rejection rates appear to be similar for all four DGP's.

Next, we compare the out-of-sample performance of various forecasting models. We consider constant, polynomial, integrable linear (in parameters) and ERP forecasting equations:

$$\widehat{y}_t(\mu, 0) = \mu, \tag{6.15}$$

$$\widehat{y}_t(\mu, \theta) = \mu + \sum_{i=1}^{p} \theta_i z_{t-1}^i, \tag{6.16}$$

$$\widehat{y}_t(\mu, \theta) = \mu + \phi(z_{t-1}) \sum_{i=1}^{p} \theta_i |z_{t-1}|^{i-1}, \tag{6.17}$$

$$\widehat{y}_t(\mu,\theta) = \mu + \phi(z_{t-1})\frac{\theta_1 + \theta_2 z_{t-1} + ... + \theta_{p+1} z_{t-1}^p}{1 + \theta_{p+2} z_{t-1} + ... + \theta_{2p+1} z_{t-1}^p}. \quad (6.18)$$

We use the absolute value in the definition of (6.17), so that all base functions have a non-zero Lebesgue measure. First, we simulate 200 observations and use observations $\{1, ..., 100\}$ to estimate the parameters in (6.15)-(6.18) (in this case, $r = 1/2$). The parameters are estimated by LS (nonlinear LS, in the case of equation (6.18)). In the second step, we construct one period ahead forecasts for observations$\{101, ..., 200\}$ using the estimated versions of (6.15)-(6.18). Finally, using predicted values of y_t, we compute the out-of-sample MSE's for the four forecasting equations.

Table 2 reports the proportion of cases in which forecasting functions (6.16)-(6.18) have smaller MSE's than that of the historic average (model (6.15)). The number of simulations is 1,000, $a \in \{1, 10\}$, and $p \in \{1, 2, 3, 4\}$. Consider the case of f_1 and $p = 1$. The numbers corresponding to the polynomial forecasting model show that the historic average provides better out-of-sample fit than predictive regression in approximately 66%-69% of the repetitions. Increasing the value of a from 1 to 10 leads only to a marginal improvement in the performance of the polynomials. Thus, by ignoring the information contained in the predictor, with high probability one can obtain better forecasts, despite the fact that F_n indicates predictive power. Performance of the integrable forecasting functions depends on the strength of the signal coming from the nonlinear component. In the case of $a = 1$, the historic average provides better out-of-sample fit in 42%-50% of the repetitions. However, in the case of $a = 10$, the integrable forecasting models perform better than the baseline model in 78%-85% of the repetitions. Similar patterns are observed for the DGPs f_2, f_3 and f_4. Hence, the result of Theorem 4.2(c) holds in finite samples, provided that the signal-to-noise ratio is sufficiently large. Table 2 shows that the performance of (6.16)-(6.18) may deteriorate as p increases. This can be explained by the slow rates of convergence in the case of integrable transformations of I(1) processes. Evidently, when $p > 1$, larger sample sizes are required in order to obtain better approximations.

Finally, we evaluate finite sample size and power properties of the in-sample (T_n) and out-of-sample (S_n) tests proposed in Section 5. We use values of $a \in \{0, 0.5, 1, 2\}$, and the nominal size of 5%. For the out-of-sample test, We consider values of $r \in \{1/4, 1/2, 3/4\}$. Critical values were calculated from 1,000 simulation repetitions. The test statistics are constructed using the estimates of θ in (6.17) for $p = 1$. We set the number of observations equal to 500, and the number of simulations to 1,000.

Table 3 reports the results. First, consider $a = 0$, which corresponds to the case of no predictive power. As one can see from the table, the actual rejection rates are close to the nominal, especially when $p = 1$. We conclude that, under the null, the asymptotic distributions provide reasonable approximations to the actual distribution of S_n and T_n in finite samples. Next, the results for $a > 0$ show that the tests have power for all choices of f. The in-sample test appears to be somewhat more powerful than the out-of-sample test in the cases of f_1, f_2 and f_4, and considerable more powerful in the case of f_3 (these results coincide with the conclusion reached by Inoue and Kilian (2004)). For f_3 and $a = 2$, in-sample test attains 75% rejection rate, while the rejection rates for the out-of-sample test are in the range 37%-48%. In the case of f_3, observed rejection rates in general are lower than for the other three DGPs.

7 Empirical example

The dividend-price ratio (dividend yield) has received much attention in the literature as a potential predictor for stock returns. In a recent study, Lewellen (2004) considered the regression of stock returns on the natural log of the dividend yield and reported strong predictive power. Goyal and Welch (2003) approached the same problem from a different perspective. They focused on out-of-sample fit and arrived at an opposite conclusion. This section evaluates the predictive power of the natural log of the dividend-price ratio (LDP) in view of the theoretical findings of the previous sections. We consider the same data as Goyal and Welch (2003): monthly observations, for the period 1946-2000, of value-weighted NYSE stock returns.

Table 4 shows the results of unit root tests for the LDP. We consider two alternative autoregressive specifications for the LDP: with and without linear deterministic trend. The first line of the table shows that the estimated autoregressive coefficient is very close to unity in both cases. Furthermore, the Phillips-Perron Z_t test is unable to reject the null hypothesis of I(1) for either specification. Finally, the strongest evidence in support of the I(1) hypothesis for the LDP comes from KPSS tests (see Kwiatkowski et al. (1992)). The KPSS test assumes stationarity under the null hypothesis. Rejection of the null suggests that there is strong evidence in favor of the nonstationary alternative. As one can see from Table 4, this is the case with the LDP. At 1% significance level, the null hypothesis is rejected for both models, with or without the deterministic trend component. We conclude that a unit root model

is a reasonable approximation for the LDP.

Next, we look at in-sample predictability. We compare three testing procedures. The first procedure is based on the usual predictive (OLS) regression of stock returns on the first lag of the LDP. The second test is based on the Fully Modified OLS (FM-OLS) estimator of the regression slope, which is corrected for endogeneity of errors. While the OLS based t-test is invalid if the errors and predictor are correlated, the FM-OLS t-statistic has a mixed normal distribution regardless of correlations between errors and the regressor (see Phillips and Hansen (1990)). It is important to emphsize that the linear model is not regarded as a true DGP, but rather as a misspecified forecasting equation. According to the results of Section 3, one should expect to see a small estimated slope coefficient and a large t-statistic if the true DGP is well approximated by the one described in Assumptions 2.1-2.3. Finally, we consider the in-sample testing approach proposed in Section 5.1. For that purpose, we use standard normal density transformation of the LDP ($p = 1$). In this case, one should reject the null of no predictive power if $|T_n|^{1/2} > z_{1-\alpha}$, where z_α is the α-quantile of the standard normal distribution. Table 5 reports the results of the tests. First, the OLS and FM-OLS estimates are of small magnitude. Further, the OLS based t-statistic is large; however, it is not significant at 5% significance level. The statistic based on FM-OLS estimates and $|T_n|^{1/2}$ statistic are both significant. Hence, in-sample evidence indicates possible predictive power of the LDP.

The large values of t-statistics should not be taken as an evidence in favor of a linear relationship between stock returns and the lagged value of the LDP. As we argue in Section 3, in this framework, it rather points in the direction of a possible nonlinear dependence.

Lastly, we compare the out-of-sample performance of the four predictive models given in equations (6.15)-(6.18). We set $p = 1$ in (6.16)-(6.18). Equation (6.15) corresponds to the assumption that the stock returns cannot be predicted from the historic values of the LDP and produces constant forecasts. Equation (6.16) is the usual predictive regression model. Equation (6.17) corresponds to the integrable forecasting function linear in the parameters. Finally, equation (6.18) describes the ERP forecasting model. For the purpose of this exercise, we select $r \in (0, 1)$ and divide the sample into two parts: observations $\{1, ..., [nr]\}$ and observations $\{[nr] + 1, ..., n\}$. In the first step, we use observations $\{1, ..., [nr]\}$ to estimate the unknown parameters in (6.15)-(6.18). In the case of (6.15), the average of y_t in the training sample is used

to estimate μ. The historic average has been used to estimate μ in (6.17) and (6.18) as well. As it was mentioned above, replacing the LS estimator of the intercept by the simple average of y_t's in the training sample has no effect on the behavior of the loss function asymptotically. However, it appears to perform better in practice. In the case of ERP, parameters are estimated by nonlinear LS. we used zeros as starting values for $\widehat{\theta}_1$, $\widehat{\theta}_2$ and $\widehat{\theta}_3$ for numerical optimization. This choice follows from the fact that $\theta_1 = \theta_2 = 0$ in (6.18) implies that the LDP has no predictive power. In the second step, estimated versions of (6.15)-(6.18) and the actual values of LDP in the forecasting sample are used to construct one period ahead forecasts for observations $\{[nr] + 1, ..., n\}$.

Table 6 reports the out-of-sample square-root MSE's, test statistic S_n defined in Section 5.2 and its p-value. The out-of-sample S_n statistic was calculated using the standard normal density transformation of the predictor. Its p-value was calculated using 10,000 simulations. we consider $r \in \{1/4, 1/2, 3/4\}$. It appears that the linear regression is the worst performer, since it is dominated by the historic average and integrable functions for all considered choices of r. The MSE's of the integrable functions is smaller than that of the historic average at all forecasting horizons. However, S_n statistic has large p-values of 34%-45%, implying that the difference between the forecasting accuracy of the historic average and integrable function (linear in parameters) is not statistically significant. Nevertheless, we find it interesting that, in the above results, applying an integrable transformation produces automatic improvement in the out-of-sample performance against linear regression and historic average. It is possible that, at the considered sample size, the out-of-sample statistic does not have enough power to reject the null of no predictability.

It is important to emphasize that the above exercise is not an attempt to obtain a best nonlinear predictor. For that purpose, the search process among integrable functions must be taken into account in the testing procedure to avoid data snooping bias. This problem can be approached along the lines described by White (2000). Development of such and ERP based tests is a part of the ongoing research project.

We offer the following interpretation of the results reported in Table 6. It is reasonable to assume that the LDP contains an autoregressive unit root. Consequently, the LDP cannot be a good predictor for stock returns during the periods which exhibit apparent trending behavior. Integrable transformations of the LDP improve out-of-sample fit because they filter out large values of the LDP. This allows one to

ignore the LDP during the trending periods and extract useful information during the other times. This interpretation is consistent with the idea that stock returns are predictable only on rare occasions, when the predictor does not show clear patterns immediately observable by all market participants.

8 Concluding remarks

In this paper, we consider forecasting of time series which contain a nonstationary, nonlinear component. The nonlinear component is modeled as an integrable transformation of the predictor, which is assumed to be an I(1) variable. It is assumed that the true form of the nonlinear component is unknown to the econometrician, and that he is forced to use some approximating functions. We show that standard tools such as t-type tests and linear regressions lead to spurious forecasts. The diagnostic tests tend to indicate predictive ability, while the forecasts based on the usual linear regression perform worse in terms of the MSE than constant forecasts, which ignore the information contained in the predictor. This paper provides general approximating results, which allow one to improve the forecast accuracy with properly chosen nonlinear forecasting functions. It is shown that one can obtain non-trivial improvements in forecast accuracy over polynomials and historic averages by using square integrable forecasting functions. Only the case of a single predictor is considered in this paper. The analogous results can be obtained in the case of multiple covariates with additively separate regression functions.

The empirical application considered in this paper is concerned with forecasting the NYSE stock returns using the dividend-price ratio. We show that integrable transformations of the dividend-price ratio provide somewhat better out-of-sample fit than the forecasts constructed from the typical linear models. The accuracy of the forecasts is improved because nonlinear transformations filter out irrelevant information.

We would like to emphasize the importance of nonstationarity in the current context. In Section 4, we show that, in the case of a nonstationary predictor z_t, the loss function converges to the L_2-distance between the true function f and the approximating function g (multiplied by the local time process):

$$\int_{-\infty}^{\infty} \left(f(x) - g\left(x, \theta\right)\right)^2 dx.$$

In contrast, in the case of a strictly stationary and ergodic predictor z_t, the loss function converges to the L_2-distance weighted by the density of the predictor:

$$\int_{-\infty}^{\infty} \left(f(x) - g\left(x, \theta\right)\right)^2 pdf_z(x)dx,$$

where pdf_z denotes the probability density function of the predictor. In the second case, polynomials can provide a good approximation since $pdf_z(x) \to 0$ as $x \to \pm\infty$, while in the nonstationary case, approximation with polynomials is impossible due to unweighted integration over the entire real line.

One of the attractive features of the tests proposed in Section 5 is that their implementation does not require pre-tests to verify that the predictor is indeed an I(1) process, and therefore no pre-test bias will be introduced. One can easily verify that these tests have correct size even when z_t is stationary.

The results in this paper are limited to the forecasting problem. However, they can be extended to nonparametric estimation of the nonlinear component.

Appendix

Proof of Theorem 3.1. We prove part (a) first. Write $\widehat{\theta}_n = A_{1n} - A_{2n} + A_{3n}$, where

$$\begin{aligned}
A_{1n} &= \left(\sum_{t=1}^{[nr]} \underline{Z}_{t-1}\underline{Z}'_{t-1}\right)^{-1} \sum_{t=1}^{[nr]} f\left(z_{t-1}\right) Z_{t-1}, \\
A_{2n} &= \left(\sum_{t=1}^{[nr]} f\left(z_{t-1}\right)\right)\left(\sum_{t=1}^{[nr]} \underline{Z}_{t-1}\underline{Z}'_{t-1}\right)^{-1} [nr]^{-1} \sum_{t=1}^{[nr]} Z_{t-1}, \\
A_{3n} &= \left(\sum_{t=1}^{[nr]} \underline{Z}_{t-1}\underline{Z}'_{t-1}\right)^{-1} \sum_{t=1}^{[nr]} u_t \underline{Z}_{t-1}.
\end{aligned}$$

It follows from Lemma 2.1(b) and the Cramér-Wold device (see Theorem 25.5 of Davidson (1994) on page 405) that

$$n^{-1} \sum_{t=1}^{[nr]} D_n^{-1/2} \underline{Z}_{t-1}\underline{Z}'_{t-1} D_n^{-1/2} \to_d \int_0^r \underline{B}(s)\underline{B}(s)'ds. \tag{8.19}$$

Assumption 3.4 and Lemma 2.1(d) imply that

$$n^{-1/2}\sum_{t=1}^{[nr]} f\left(z_{t-1}\right) Z_{t-1} = O_p(1). \tag{8.20}$$

Therefore, it follows from (8.19) and (8.20) that

$$n^{1/2} D_n^{1/2} A_{1n} = o_p(1). \tag{8.21}$$

Next, joint convergence in Lemma 2.1(b),(d) and the CMT imply that

$$n^{1/2} D_n^{1/2} A_{2n} \to_d r^{-1}\ell(r) \int_{-\infty}^{\infty} f(x)dx \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{-1} \int_0^r B(s)ds. \tag{8.22}$$

Finally, Lemma 2.1(c) implies that

$$n^{1/2} D_n^{1/2} A_{3n} \to_d \left(\int_0^r \underline{B}(s)\underline{B}(s)'ds\right)^{-1} \int_0^r \underline{B}(s)du(s). \tag{8.23}$$

The result in part (a) follows from (8.21)-(8.23) and the joint convergence in Lemma 2.1.

The result in part (b) of the theorem follows immediately from the definition of $\widehat{\mu}_n$, Lemma 2.1 and part (a) of the theorem.

For part (c) of the theorem, it is sufficient to show that $\widehat{\sigma}_{u,n}^2$ in (3.5) converges in probability to σ_u^2. The result will follow from part (a) and the CMT. Define the averages $\overline{f}_n = [nr]^{-1}\sum_{t=1}^{[nr]} f\left(z_{t-1}\right)$, and $\overline{u}_n = [nr]^{-1}\sum_{t=1}^{[nr]} u_t$. Write

$$\sum_{t=1}^{[nr]} \left(y_t - \widehat{\mu}_n - Z_{t-1}'\widehat{\theta}_n\right)^2 = \sum_{t=1}^{[nr]} \left(\left(f\left(z_{t-1}\right) - \overline{f}_n\right) + \left(u_t - \overline{u}_n\right) - \underline{Z}_{t-1}'\widehat{\theta}_n\right)^2. \tag{8.24}$$

We have the following results:

$$\begin{aligned}\sum_{t=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)^2 &= \sum_{t=1}^{[nr]} f^2\left(z_{t-1}\right) - \left([nr]^{-1/2}\sum_{i=1}^{[nr]} f\left(z_{t-1}\right)\right)^2 \\ &= O_p\left(n^{1/2}\right), \end{aligned} \tag{8.25}$$

$$\widehat{\theta}_n' \sum_{t=1}^{[nr]} \underline{Z}_{t-1} \underline{Z}_{t-1}' \widehat{\theta}_n = O_p(1), \tag{8.26}$$

$$\sum_{t=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)\left(u_t - \overline{u}_n\right) = O_p\left(n^{1/4}\right), \tag{8.27}$$

$$\sum_{t=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right) \underline{Z}_{t-1}' \widehat{\theta}_n = o_p\left(1\right), \tag{8.28}$$

$$\sum_{t=1}^{[nr]} \left(u_t - \overline{u}_n\right) \underline{Z}_{t-1}' \widehat{\theta}_n = O_p\left(1\right). \tag{8.29}$$

Equation (8.25) follows from Assumption 2.3 and Lemma 2.1(d); (8.26) is implied by part (a) of the theorem and Lemma 2.1(b); (8.27) is due to Lemma 2.1(e); (8.28) follows from part (a) of the theorem, Lemma 2.1(d) and Assumption 3.4; and, lastly, (8.29) follows from 2.1(c) and part (a) of the theorem. Next, (8.24)-(8.29) together imply that

$$\sum_{t=1}^{[nr]} \left(y_t - \widehat{\mu}_n - Z_{t-1}' \widehat{\theta}_n\right)^2 = \sum_{t=1}^{[nr]} u_t^2 + \sum_{t=1}^{[nr]} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)^2 + O_p\left(n^{1/4}\right).$$

The result follows from Assumption 2.2(a) and (b) and the MDS LLN (see for, example, Theorem 20.11 on page 315 of Davidson (1994)).

The proof of part (d) is similar to the derivation of the probability limit of s_n^2 in (b). Using the same arguments as in (8.24)-(8.29), one can write $nQ_n\left(\widehat{\mu}_n, \widehat{\theta}_n\right)$ as follows:

$$\begin{aligned}
& \sum_{t=[nr]+1}^{n} \left(f\left(z_{t-1}\right) - \overline{f}_n\right)^2 + O_p\left(n^{1/4}\right) \\
= & \sum_{t=[nr]+1}^{n} f^2\left(z_{t-1}\right) + \frac{n-[nr]}{[nr]^2}\left(\sum_{t=1}^{[nr]} f\left(z_{t-1}\right)\right)^2 \\
& - \frac{1}{[nr]} \sum_{t=1}^{[nr]} f\left(z_{t-1}\right) \sum_{t=[nr]+1}^{n} f\left(z_{t-1}\right) + O_p\left(n^{1/4}\right) \\
= & \sum_{t=[nr]+1}^{n} f^2\left(z_{t-1}\right) + O_p\left(n^{1/4}\right).
\end{aligned} \tag{8.30}$$

The result follows from (8.30), Assumption 2.3 and Lemma 2.1(d).

We prove part (e) of the theorem now. Write

$$\begin{aligned}
&(n-[nr])\left(Q_n\left(\widehat{\mu}_n,\widehat{\theta}_n\right)-Q_n\left(\widehat{\mu}_{0,n},0\right)\right)\\
=&\sum_{t=[nr]+1}^{n}\left(y_t-\widehat{\mu}_{0,n}-\underline{Z}_{t-1}'\widehat{\theta}_n\right)^2-\sum_{t=[nr]+1}^{n}\left(y_t-\widehat{\mu}_{0,n}\right)^2\\
=&B_{1n}-2B_{2n}-2B_{3n},
\end{aligned}$$

where

$$\begin{aligned}
B_{1n}&=\widehat{\theta}_n'\sum_{t=[nr]+1}^{n}\underline{Z}_{t-1}\underline{Z}_{t-1}'\widehat{\theta}_n,\\
B_{2n}&=\sum_{t=[nr]+1}^{n}\left(f\left(z_{t-1}\right)-\overline{f}_n\right)\underline{Z}_{t-1}'\widehat{\theta}_n,\\
B_{3n}&=\sum_{t=[nr]+1}^{n}\left(u_t-\overline{u}_n\right)\underline{Z}_{t-1}'\widehat{\theta}_n.
\end{aligned}$$

Due to the result in part (a),

$$B_{1n}\ \rightarrow\ {}_d\Psi'\left(\int_r^1\underline{B}(s)\underline{B}(s)'ds\right)\Psi, \tag{8.31}$$

$$\begin{aligned}
B_{2n}=&-\left[nr\right]^{-1}\sum_{t=1}^{[nr]}f\left(z_{t-1}\right)\sum_{t=[nr]+1}^{n}\underline{Z}_{t-1}'\widehat{\theta}_n\\
&-\sum_{t=[nr]+1}^{n}f\left(z_{t-1}\right)\left[nr\right]^{-1}\sum_{t=1}^{[nr]}Z_{t-1}'\widehat{\theta}_n+\sum_{t=[nr]+1}^{n}f\left(z_{t-1}\right)Z_{t-1}'\widehat{\theta}_n,
\end{aligned}$$

where the last term is $o_p(1)$. We have

$$\begin{aligned}
B_{2n}\ \rightarrow\ &{}_d-r^{-1}\left(\int_{-\infty}^{\infty}f(x)dx\right)\\
&\times\left(\ell(r)\int_r^1\underline{B}(s)d(s)+\int_1^r d\ell(s)\int_0^r B(s)ds\right)'\Psi, \qquad (8.32)\\
B_{3n}\ \rightarrow\ &{}_d\left(\int_r^1\underline{B}(s)du(s)-r^{-1}u(r)\int_r^1B(s)ds\right)'\Psi. \qquad (8.33)
\end{aligned}$$

It follows then from (8.31)-(8.33) that the limiting random variable $(1-r)\Delta$ is given by

$$\begin{aligned}
&\Psi'\left(\int_r^1 \underline{B}(s)\underline{B}(s)'ds\right)\Psi \\
&+2r^{-1}\left(\int_{-\infty}^{\infty} f(x)dx\right)\Psi'\left(\ell(r)\int_r^1 \underline{B}(s)d(s)+\int_r^1 d\ell(s)\int_0^r B(s)ds\right) \\
&-2\Psi'\left(\int_r^1 \underline{B}(s)du(s)-r^{-1}u(r)\int_r^1 B(s)d(s)\right).
\end{aligned}$$

The support of Δ includes both negative and positive parts of the real line since the following matrix is indefinite:

$$\begin{pmatrix} \int_r^1 \underline{B}(s)\underline{B}(s)'ds & I_p \\ I_p & 0 \end{pmatrix},$$

and Δ is a quadratic form involving the above matrix. □

The following lemma is used in the proof of Theorem **4.2**. It considers the case of extremum estimators with multiple optimal points. Let "$\Rightarrow$" denote the weak convergence of stochastic processes.

Lemma 8.2. Suppose that

$$\left(Q_{1n}(\theta), Q_{2n}(\theta)\right) \Longrightarrow \left(Q_1(\theta), Q_2(\theta)\right),$$

where $(Q_1(\theta), Q_2(\theta))$ is a stochastic processes indexed by $\theta \in \Theta$, and Θ is a compact subset of R^p. Define

$$\Theta^* = \left\{\theta^* \in \Theta : P\left(Q_1(\theta^*) = \inf_{\theta\in\Theta} Q_1(\theta)\right) = 1\right\}.$$

Let Θ_n be the set of values of θ that minimize $Q_{1n}(\theta)$ on Θ. Then,

(a) $\sup_{\theta_n\in\Theta_n} d(\theta_n, \Theta^*) \to_p 0$ as $n \to \infty$.

(b) Suppose that $Q_{2n}(\theta)$ is stochastically equicontinuous on Θ, and that the following condition is satisfied for all $\varepsilon > 0$

$$P\left(\sup_{\theta_1^*,\theta_2^*\in\Theta^*} |Q_2(\theta_1^*) - Q_2(\theta_2^*)| \geq \varepsilon\right) = 0. \tag{8.34}$$

Then, as $n \to \infty$

$$\sup_{\theta_n \in \Theta_n} \sup_{\theta^* \in \Theta^*} |Q_{2n}(\theta_n) - Q_{2n}(\theta^*)| \to_p 0.$$

Proof of Lemma 8.2. (a) As $n \to \infty$,

$$\overline{\lim} P\left(\sup_{\theta_n \in \Theta_n} d(\theta_n, \Theta^*) \geq \delta\right) \leq \overline{\lim} P\left(\inf_{\theta \in \Theta: d(\theta, \Theta^*) \geq \delta} Q_{1n}(\theta) \leq \inf_{\theta^* \in \Theta^*} Q_{1n}(\theta^*)\right).$$

Define

$$h(Q) = 1\left\{\inf_{\theta \in \Theta: d(\theta, \Theta^*) \geq \delta} Q(\theta) \leq \inf_{\theta^* \in \Theta^*} Q(\theta^*)\right\}.$$

The definition of Θ^* implies that $h(Q_1) = 0$ with probability one. Next, for all $\varepsilon > 0$,

$$\begin{aligned} \overline{\lim} P\left(\inf_{\theta \in \Theta: d(\theta, \Theta^*) \geq \delta} Q_{1n}(\theta) \leq \inf_{\theta^* \in \Theta^*} Q_{1n}(\theta^*)\right) &= \overline{\lim} P(h(Q_{1n}) \geq \varepsilon) \\ &\leq P(h(Q_1) \geq \varepsilon) \\ &= 0. \end{aligned} \tag{8.35}$$

The inequality in (8.35) follows from weak convergence and the continuous mapping theorem (CMT) (see Theorem 22.11 of Davidson (1994) on page 355).

(b) As $n \to \infty$,

$$\begin{aligned} & \overline{\lim} P\left(\sup_{\theta_n \in \Theta_n} \sup_{\theta^* \in \Theta^*} |Q_{2n}(\theta_n) - Q_{2n}(\theta^*)| \geq \varepsilon\right) \\ \leq\ & \overline{\lim} P\left(\sup_{\theta_n \in \Theta_n} \sup_{\theta^* \in \Theta^*} |Q_{2n}(\theta_n) - Q_{2n}(\theta^*)| \geq \varepsilon,\ \sup_{\theta_n \in \Theta_n} d(\theta_n, \Theta^*) < \delta\right) \\ & + \overline{\lim} P\left(\sup_{\theta_n \in \Theta_n} d(\theta_n, \Theta^*) \geq \delta\right). \end{aligned} \tag{8.36}$$

Next, the condition $\sup_{\theta_n \in \Theta_n} d(\theta_n, \Theta^*) < \delta$ implies that for all $\theta_n \in \Theta_n$ there exists $\theta_n^* \in \Theta^*$ such that $\|\theta_n^* - \theta_n\| < \delta$. The first summand on the right-hand side of (8.36) is bounded by

$$\begin{aligned} & \overline{\lim} P\left(\sup_{\theta_n \in \Theta_n} |Q_{2n}(\theta_n) - Q_{2n}(\theta_n^*)| + \sup_{\theta^* \in \Theta^*} |Q_{2n}(\theta_n^*) - Q_{2n}(\theta^*)| \geq \varepsilon\right) \\ \leq\ & \overline{\lim} P\left(\sup_{\theta^* \in \Theta^*} \sup_{\|\theta^* - \theta\| < \delta} |Q_{2n}(\theta) - Q_{2n}(\theta^*)|\right. \end{aligned}$$

$$+ \sup_{\theta_1^*, \theta_2^* \in \Theta^*} |Q_{2n}(\theta_1^*) - Q_{2n}(\theta_2^*)| \geq \varepsilon \Bigg).$$

Now, weak convergence, the CMT and (8.34) imply that

$$\sup_{\theta_1^*, \theta_2^* \in \Theta^*} |Q_{2n}(\theta_1^*) - Q_{2n}(\theta_2^*)| \to_p 0.$$

The second summand on the right-hand side of (8.36) is $o(1)$ due to part (a) of the lemma. The desired result follows from the stochastic equicontinuity of Q_{2n} and Slutsky's theorem (see Theorem 18.10(ii) of Davidson (1994) on page 286). □

Proof of Theorem 4.2. Concentrating out μ as in (4.11), write the in-sample MSE as

$$\begin{aligned} n^{1/2} MSE_n(\theta) &= n^{-1/2} \sum_{t=1}^{[nr]} \left(y_t - \widehat{\mu}_{0,n} - g(z_{t-1}, \theta) + \overline{g}_n(\theta) \right)^2 \\ &= n^{-1/2} \sum_{t=1}^{[nr]} \left(f(z_{t-1}) - g(z_{t-1}, \theta) \right)^2 + n^{-1/2} \sum_{t=1}^{[nr]} (u_t - \overline{u}_n)^2 \\ &\quad + R_n(\theta), \end{aligned}$$

where $R_n(\theta) = R_{1n}(\theta) + 2R_{2n}(\theta) + 2R_{3n}(\theta) + 2R_{4n}(\theta)$, and

$$\begin{aligned} R_{1n}(\theta) &= n^{1/2} \left(\overline{f}_n - \overline{g}_n(\theta) \right)^2, \\ R_{2n}(\theta) &= \left(\overline{f}_n - \overline{g}_n(\theta) \right) n^{-1/2} \sum_{t=1}^{[nr]} \left(f(z_{t-1}) - g(z_{t-1}, \theta) \right), \\ R_{3n}(\theta) &= n^{-1/2} \sum_{t=1}^{[nr]} \left(f(z_{t-1}) - g(z_{t-1}, \theta) \right) (u_t - \overline{u}_n), \\ R_{4n}(\theta) &= \left(\overline{f}_n - \overline{g}_n(\theta) \right) n^{-1/2} \sum_{t=1}^{[nr]} (u_t - \overline{u}_n). \end{aligned}$$

Next, we show that each of the components of $R_n(\theta)$ is $o_p(1)$ uniformly in θ. For $R_{1n}(\theta)$, write $R_{1n}(\theta) = n^{-1/2} \left(n^{-1/2} \sum_{t=1}^{[nr]} \left(f(z_{t-1}) - g(z_{t-1}, \theta) \right) \right)^2$. Then, Assumptions 2.3 and 4.5, uniform convergence in Theorem 3.2 of Park and Phillips (2001) and the CMT imply that $\sup_{\theta \in \Theta} |R_{1n}(\theta)| = o_p(1)$. In a similar way, one can show

that $R_{2n}(\theta)$ and $R_{4n}(\theta)$ converge to zero in probability uniformly in θ. Finally, $\sup_{\theta\in\Theta} R_{3n}(\theta) = o_p(1)$ as it follows from Lemma A7(b) of Park and Phillips (2001). It follows now from Assumptions 2.3, 4.5 and Theorem 3.2 of Park and Phillips (2001) that

$$n^{1/2} MSE_n(\theta) - n^{-1/2} \sum_{t=1}^{[nr]} (u_t - \overline{u}_n)^2 \Longrightarrow \ell(r) \int_{-\infty}^{\infty} (f(x) - g(x,\theta))^2 \, dx. \tag{8.37}$$

The result in (8.37) and Lemma 8.2(a) together imply that

$$\sup_{\theta\in\Theta_n} d(\theta_n, \Theta^*) \to_p 0, \tag{8.38}$$

which completes the proof of part (a) of the Theorem.

For part (b), write

$$\begin{aligned} \sup_{\theta_n\in\Theta_n} \left| \widehat{\mu}_n\left(\widehat{\theta}_n\right) - \mu^* \right| \quad \leq \quad & \left| [nr]^{-1} \sum_{t=1}^{[nr]} \left(f(z_{t-1}) - g(z_{t-1}, \theta^*) \right) \right| + \left| [nr]^{-1} \sum_{t=1}^{[nr]} u_t \right| \\ & + \sup_{\theta\in\Theta} \left| R_{5n}(\theta^*, \theta) \right|, \end{aligned} \tag{8.39}$$

where $\theta^* \in \Theta^*$, and

$$R_{5n}(\theta^*, \theta) = [nr]^{-1} \sum_{t=1}^{[nr]} \left(g(z_{t-1}, \theta^*) - g(z_{t-1}, \theta) \right).$$

The first two summands on the right-hand side of (8.39) are $o_p(1)$, as it follows from Lemma 2.1(a) and (d); $\sup_{\theta\in\Theta} |R_{5n}(\theta^*, \theta)| = o_p(1)$ due to the assumptions of the theorem, Theorem 3.2 of Park and Phillips (2001) and the CMT. The result of part (b) of the theorem follows.

For part (c) of the theorem, note that $Q_n\left(\widehat{\mu}_n\left(\widehat{\theta}_n\right), \widehat{\theta}_n\right)$ depends on the out-of-sample MSE. Hence, similarly to part (a), one can show that

$$n^{1/2} Q_n\left(\widehat{\mu}_n(\theta), \theta\right) \Rightarrow (1-r)^{-1} \left(\ell(1) - \ell(r)\right) \int_{-\infty}^{\infty} (f(x) - g(x,\theta))^2 \, dx.$$

Next, fix $\theta^* \in \Theta^*$. Define $R_{6,n}(\theta_n, \theta^*) = Q_n\left(\widehat{\mu}_n(\theta_n), \theta_n\right) - Q_n\left(\widehat{\mu}_n(\theta^*), \theta^*\right)$. It follows

from Lemma 8.2(b) that

$$\sup_{\theta_n \in \Theta_n, \theta^* \in \Theta^*} \left| n^{1/2} R_{6,n}\left(\widehat{\theta}_n, \theta^*\right) \right| \to_p 0.$$

Hence,

$$\begin{aligned} n^{1/2} Q_n\left(\widehat{\mu}_n\left(\widehat{\theta}_n\right), \widehat{\theta}_n\right) &= n^{1/2} Q_n\left(\widehat{\mu}_n\left(\theta^*\right), \theta^*\right) + o_p(1) \\ &\to_d (1-r)^{-1}\left(\ell(1) - \ell(r)\right) M^*, \end{aligned}$$

where the last result holds uniformly in $\widehat{\theta}_n$. □

Proof of Theorem 5.3. Under the local alternative, the re-scaled slope coefficient $n^{1/4}\widehat{\theta}_n$ is given by

$$\begin{aligned} &\left(n^{-1/2} \sum_{t=1}^{n} \left(\Phi\left(z_{t-1}\right) - \overline{\Phi}_n\right)\left(\Phi\left(z_{t-1}\right) - \overline{\Phi}_n\right)' \right)^{-1} \times \\ &\left(n^{-1/2} \sum_{t=1}^{n} f\left(z_{t-1}\right)\left(\Phi\left(z_{t-1}\right) - \overline{\Phi}_n\right) + n^{-1/4} \sum_{t=1}^{n} u_t \left(\Phi\left(z_{t-1}\right) - \overline{\Phi}_n\right) \right) = \\ &\left(n^{-1/2} \sum_{t=1}^{n} \Phi\left(z_{t-1}\right)\Phi\left(z_{t-1}\right)' \right)^{-1} \times \\ &\left(n^{-1/2} \sum_{t=1}^{n} f\left(z_{t-1}\right)\Phi\left(z_{t-1}\right) + n^{-1/4} \sum_{t=1}^{n} u_t \Phi\left(z_{t-1}\right) \right) + o_p(1). \end{aligned} \tag{8.40}$$

The equality above follows because $n^{1/2}\overline{\Phi}_n\overline{\Phi}_n'$ and $n^{-1/2}\sum_{t=1}^{n} f\left(z_{t-1}\right)\overline{\Phi}_n$ are $O_p\left(n^{-1/2}\right)$ by Assumption 2.3, 5.7 and Lemma 2.1(d); and since $n^{-1/4}\sum_{t=1}^{n} u_t\overline{\Phi}_n$ is $O_p\left(n^{-1/4}\right)$ by Assumption 5.7 and Lemma 2.1(e). The Cramér-Wold device, Lemma 2.1(d),(e) and Assumption 5.7 imply that

$$n^{-1/2} \sum_{t=1}^{n} \Phi\left(z_{t-1}\right)\Phi\left(z_{t-1}\right)' \to_d \ell(1) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx, \tag{8.41}$$

$$n^{-1/2} \sum_{t=1}^{n} f\left(z_{t-1}\right)\Phi\left(z_{t-1}\right) \to_d \ell(1) \int_{-\infty}^{\infty} f(x)\Phi(x) dx. \tag{8.42}$$

$$n^{-1/4} \sum_{t=1}^{n} u_t \Phi\left(z_{t-1}\right) \to_d \left(\sigma_u^2 \ell(1) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx\right)^{1/2} W(1), \tag{8.43}$$

where $W(r)$ is a Brownian motion independent of $\ell(r)$. In fact, convergence in (8.41)-(8.43) is joint. Hence, $n^{1/4}\widehat{\theta}_n$ converges in distribution to

$$\left(\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx\right)^{-1}\int_{-\infty}^{\infty}f(x)\Phi(x)dx$$
$$+\sigma_u\left(\ell(1)\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx\right)^{-1/2}W(1). \tag{8.44}$$

Finally, by a similar argument as in the proof of Theorem 3.1(c), $\widehat{\sigma}^2_{u,n}\to_p\sigma^2_u$. Therefore, it follows from (8.44) that

$$T_n\to_d\left\|\left(\ell(1)\int_{-\infty}^{\infty}\Phi(x)\Phi(x)'dx\right)^{-1/2}\int_{-\infty}^{\infty}\frac{f(x)\Phi(x)}{\sigma_u}dx+W(1)\right\|^2.\ \square$$

Proof of Theorem 5.4. The difference between out-of-sample error functions is

$$\begin{aligned}
&Q_n\left(\widehat{\mu}_{0,n},0\right)-Q_n\left(\widehat{\mu}_n,\widehat{\theta}_n\right)\\
=&\sum_{t=[nr]+1}^{n}\left(y_t-\widehat{\mu}_{0,n}\right)^2-\sum_{t=[nr]+1}^{n}\left(y_t-\widehat{\mu}_{0,n}-\left(\Phi\left(z_{t-1}\right)-\overline{\Phi}_{[nr]}\right)'\widehat{\theta}_n\right)^2\\
=&\sum_{t=[nr]+1}^{n}\left(n^{-1/4}f\left(z_{t-1}\right)+u_t\right)^2\\
&-\sum_{t=[nr]+1}^{n}\left(n^{-1/4}f\left(z_{t-1}\right)+u_t-\Phi\left(z_{t-1}\right)'\widehat{\theta}_n\right)^2+o_p(1)\\
=&-K_{1n}+2K_{2n}+2K_{3n}+o_p(1).
\end{aligned}\tag{8.45}$$

where

$$\begin{aligned}
K_{1n}&=\widehat{\theta}_n'\sum_{t=[nr]+1}^{n}\Phi\left(z_{t-1}\right)\Phi\left(z_{t-1}\right)'\widehat{\theta}_n,\\
K_{2n}&=n^{-1/4}\sum_{t=[nr]+1}^{n}f\left(z_{t-1}\right)\Phi\left(z_{t-1}\right)'\widehat{\theta}_n,\\
K_{3n}&=\sum_{t=[nr]+1}^{n}u_t\Phi\left(z_{t-1}\right)'\widehat{\theta}_n.
\end{aligned}$$

The second equality in (8.45) follows by the argument similar to that in the proof of Theorem 5.3, equation (8.40).

Next, let a_1 and a_2 be two constants, and W be a $N(0, I_p)$ random vector. By the same argument as in the proof of the second part of Lemma 1 of Chang and Park (2003), we have that

$$\begin{aligned} & a_1 n^{-1/4} \sum_{t=1}^{[nr]} u_t \Phi(z_{t-1}) + a_2 n^{-1/4} \sum_{t=[nr]+1}^{n} u_t \Phi(z_{t-1}) \\ \to \;& {}_d\sigma_u \left(a_1^2 \ell(r) + a_2^2 (\ell(1) - \ell(r)) \right)^{1/2} \left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1/2} W. \end{aligned}$$

Then, the Cramér-Wold device implies that

$$\begin{aligned} & n^{-1/4} \sum_{t=1}^{[nr]} u_t \Phi(z_{t-1}) \\ \to \;& {}_d\sigma_u \left(\ell(r) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1/2} W_1, \qquad (8.46) \\ & n^{-1/4} \sum_{t=[nr]+1}^{n} u_t \Phi(z_{t-1}) \\ \to \;& {}_d\sigma_u \left((\ell(1) - \ell(r)) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1/2} W_2, \qquad (8.47) \end{aligned}$$

where W_1, W_2 are independent $N(0, I_p)$ random vectors, and convergence is joint. The following results follow from Assumptions 2.3, 5.7, Lemma 2.1(d), equations (8.46) and (8.47), and hold jointly:

$$\begin{aligned} K_{1n} \to_d \; & (\ell(1) - \ell(r)) \left\| \left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1/2} \int_{-\infty}^{\infty} f(x)\Phi(x) dx \right. \\ & \left. + \sigma_u \ell^{-1/2}(r) W_1 \right\|, \qquad (8.48) \\ K_{2n} \to_d \; & (\ell(1) - \ell(r)) \int_{-\infty}^{\infty} f(x)\Phi(x)' dx \\ & \times \left(\left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)' dx \right)^{-1} \int_{-\infty}^{\infty} f(x)\Phi(x) dx \right. \end{aligned}$$

$$+\sigma_u \left(\ell(r) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1/2} W_2 \Bigg), \tag{8.49}$$

$$K_{3n} \rightarrow_d \sigma_u W_2' \left((\ell(1) - \ell(r)) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{1/2}$$
$$\times \left(\left(\int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1} \int_{-\infty}^{\infty} f(x)\Phi(x)dx \right.$$
$$\left. +\sigma_u \left(\ell(r) \int_{-\infty}^{\infty} \Phi(x)\Phi(x)'dx \right)^{-1/2} W_1 \right). \tag{8.50}$$

The desired result follows from equations (8.45) and (8.48)-(8.50). □

References

Campbell, J. Y., Thompson, S. B., 2004. Predicting the equity premium out of sample: Can anything beat the historical average?, NBER Working Paper 11468.

Chang, Y., Park, J. Y., 2003. Index models with integrated time series. Journal of Econometrics 114, 73–106.

Chang, Y., Park, J. Y., Phillips, P. C. B., 2001. Nonlinear econometric models with cointegrated and deterministically trending regressors. Econometrics Journal 4, 1–36.

Chung, K., Williams, R., 1990. Introduction to Stochastic Integration, 2nd Edition. Birkhäuser, Boston.

Cochrane, J. H., 1997. Where is the market going? Uncertain facts and novel theories. Economic Perspectives 21, 3–37.

Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press, New York.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13 (3), 253–263.

Fama, E. F., 1991. Efficient capital markets: II. Journal of Finance 46, 1575–1617.

Goyal, A., Welch, I., 2003. A note on predicting returns with financial ratios, working paper, Yale University.

Goyal, A., Welch, I., 2004. A comprehensive look at the empirical performance of equity premium prediction, NBER Working Paper 10483.

Granger, C. W. J., Teräsvirta, T., 1993. Modelling Nonlinear Economic Relationships. Oxford University Press, New York.

Hong, S. H., Phillips, P. C. B., 2005. Testing linearity in cointegrating relations with an application to purchasing power parity, Cowles Foundation Discussion Paper 1541.

Hu, L., Phillips, P. C. B., 2004. Nonstationary discrete choice. Journal of Econometrics, 103–138.

Inoue, A., Kilian, L., 2004. In-sample or out-of-sample tests of predictability: Which one should we use? Econometric Reviews 23, 371–402.

Jeganathan, P., 2003. Second order limits of functionals of sums of linear processes that converge to fractional stable motions, working paper, Indian Statistical Institute.

Kasparis, I., 2004. Functional form misspecification in regression with integrated time series. Ph.D. thesis, University of Southampton, Southampton, UK.

Kilian, L., Taylor, M. P., 2003. Why is it so difficult to beat the random walk forecast of exchange rates? Journal of International Economics 60 (1).

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics 54, 159–178.

Lewellen, J., 2004. Predicting returns with financial ratios. Journal of Financial Economics 74, 209–235.

Park, J. Y., Phillips, P. C. B., 1999. Asymptotics for nonlinear transformations of integrated time series. Econometric Theory 15, 269–298.

Park, J. Y., Phillips, P. C. B., 2001. Nonlinear regressions with integrated time series. Econometrica 69, 117–161.

Phillips, P. C. B., 1983. Best uniform and modified pade approximants to probability densities in econometrics. In: Hildenbrand, W. (Ed.), Advances in Econometrics. Cambridge University Press, pp. 123–167.

Phillips, P. C. B., Hansen, B. E., 1990. Statistical inference in instrumental variables regression with I(1) processes. Review of Economic Studies 57, 99–125.

Phillips, P. C. B., Jin, S., Hu, L., 2007. Nonstationary discrete choice: A corrigendum and addendum. Journal of Econometrics, forthcoming.

West, K. D., McCracken, M. W., 1998. Regression-based tests of predictive ability. International Economic Review 39, 817–840.

White, H., 2000. A reality check for data snooping. Econometrica 68 (5), 1097–1126.

Figure 1: Simulated sample path of $y_t = f(z_{t-1}) + u_t$, where $f(x) = 10\phi(x)$, and ϕ is the standard normal density function

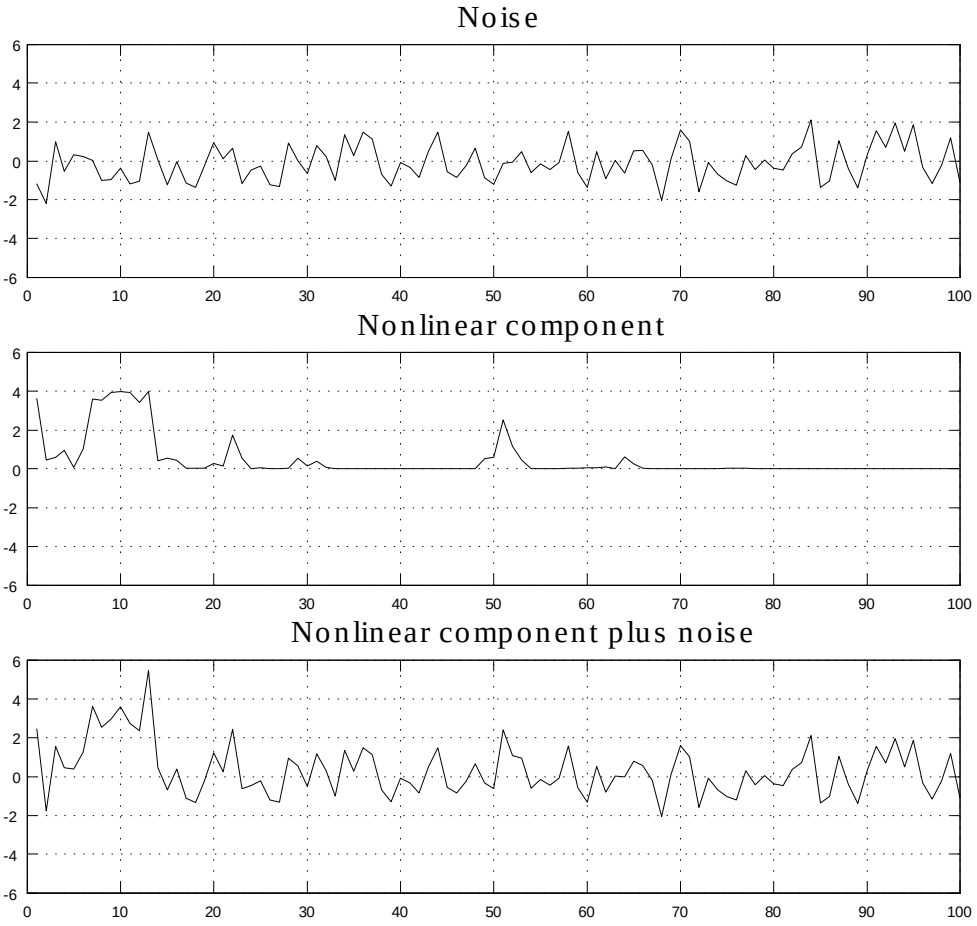

Table 1: Simulated rejection rates of F_n test for 0.05 significance level, different DGP's (f), values of the order of approximation (p) and signal-to-noise ratio (a)

p	f_1	f_2	f_3	f_4
			$a = 1$	
1	0.115	0.113	0.112	0.103
2	0.126	0.108	0.121	0.121
3	0.142	0.136	0.136	0.133
4	0.176	0.165	0.157	0.136
			$a = 2$	
1	0.233	0.205	0.247	0.264
2	0.295	0.256	0.314	0.299
3	0.340	0.303	0.353	0.355
4	0.419	0.376	0.433	0.398
			$a = 4$	
1	0.411	0.355	0.459	0.556
2	0.536	0.457	0.603	0.782
3	0.594	0.556	0.665	0.878
4	0.691	0.629	0.748	0.906
			$a = 8$	
1	0.538	0.464	0.600	0.654
2	0.707	0.597	0.795	0.908
3	0.758	0.695	0.825	0.967
4	0.827	0.761	0.885	0.981

Table 2: Proportion of simulation repetitions where the out-of-sample MSE of the corresponding model is smaller than that of the historic average, for different DGPs (f), values of the order of approximation (p) and signal-to-noise ratio (a)

		integrable			integrable	
p	polynomial	linear	ERP	polynomial	linear	ERP
		f_1, $a = 1$			f_1, $a = 10$	
1	0.299	0.581	0.503	0.342	0.846	0.776
2	0.205	0.539	0.462	0.257	0.843	0.784
3	0.107	0.494	0.454	0.175	0.818	0.780
4	0.089	0.426	0.486	0.154	0.809	0.793
		f_2, $a = 1$			f_2, $a = 10$	
1	0.292	0.574	0.618	0.357	0.867	0.821
2	0.202	0.511	0.540	0.271	0.850	0.805
3	0.103	0.452	0.523	0.186	0.831	0.819
4	0.087	0.403	0.508	0.177	0.838	0.819
		f_3, $a = 1$			f_3, $a = 10$	
1	0.275	0.511	0.607	0.357	0.815	0.877
2	0.216	0.518	0.525	0.254	0.759	0.819
3	0.108	0.450	0.502	0.183	0.770	0.844
4	0.085	0.406	0.487	0.174	0.758	0.865
		f_4, $a = 1$			f_4, $a = 10$	
1	0.174	0.587	0.531	0.281	0.919	0.726
2	0.122	0.513	0.462	0.226	0.923	0.574
3	0.066	0.469	0.458	0.180	0.925	0.646
4	0.063	0.415	0.449	0.172	0.928	0.681

Table 3: Simulated size and power of in-sample T_n and out-of-sample S_n tests for the order of approximation $p = 1$, different DGPs (f), values of the signal-to-noise ratio (a), and 0.05 significance level

		out-of-sample S_n		
a	in-sample T_n	$r = 1/4$	$r = 1/2$	$r = 3/4$
0	0.050	0.049	0.051	0.056
		f_1		
0.5	0.192	0.152	0.188	0.193
1	0.571	0.447	0.488	0.534
2	0.993	0.903	0.934	0.958
		f_2		
0.5	0.154	0.134	0.161	0.147
1	0.454	0.357	0.389	0.374
2	0.969	0.821	0.845	0.809
		f_3		
0.5	0.102	0.088	0.099	0.085
1	0.268	0.191	0.212	0.169
2	0.753	0.483	0.450	0.369
		f_4		
0.5	0.178	0.147	0.179	0.163
1	0.541	0.424	0.475	0.399
2	0.990	0.875	0.915	0.773

Table 4: Unit root tests for the log of the dividend-price ratio.

	intercept	intercept and trend
autoregressive coefficient	0.997	0.990
Phillips-Perron test		
Z_t statistic	−0.702	−1.808
10% critical value	−2.569	−3.132
KPSS test		
statistic	2.969	0.500
1% critical value	0.739	0.216

Table 5: In-sample performance of the log of the dividend-price ratio

	regression slope estimate	t-statistic
OLS	0.0092	1.92
FM-OLS	0.0065	4.37
Integrable	1.7442	2.61

Table 6: Out-of-Sample Root MSE $\times 10^2$, S_n test statistic and its estimated p-value for different relative sizes of the training sample (r)

r	1/4	1/2	3/4
historic average	4.2223	4.4500	4.0765
linear regression	4.2812	4.4744	4.2521
integrable linear	4.2119	4.4232	4.0692
integrable ERP	4.2046	4.4235	4.0692
S_n statistic	0.7718	5.8269	0.5919
p-value of S_n	0.3419	0.3526	0.4519